DOCUMENT RESUME

ED 046 042 CG 006 133

AUTHOR	Engel, John
TITLE	An Approach to Standardizing Human Performance Assessment.
INSTITUTION	Human Resources Research Organization, Alexandria, Va.; Texas Tech Univ., Lubbock.
PUB DATE	Mar 70
NOTE	10p.; Presentation at the Planning Conference of Standardization of Tasks and Measures for Human Factors Research, Lubbock, Texas, March, 1970
EDRS PRICE	EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS	*Classification, Criteria, Evaluation Methods, *Evaluation Needs, *Human Development, Job Analysis, Measurement Techniques, *Performance Criteria, Performance Factors, *Task Analysis, Task Performance

ABSTRACT

The standardization and evaluation of methods of performance assessment represents an important area of concern. In this paper an approach that concentrates on two critical areas and the relationship between them is discussed. These are: (1) a task classification system; and (2) a performance measure classification system. An example is presented that illustrates some preliminary research related to the use of a performance measure classification system. The paper concludes by suggesting areas and directions for future research efforts. (Author)

Professional Paper 26.70
October 1970

An Approach to Standardizing Human Performance Assessment

by

John D. Engel

Presentation at the
Planning Conference of 'Standardization of Tasks and Measures for Human Factors Research'
Texas Technological University
Lubbock, Texas March 1970

HumRRO

HUMAN RESOURCES RESEARCH ORGANIZATION

The Human Resources Research Organization (HumRRO) is a nonprofit corporation established in 1969 to conduct research in the field of training and education. It is a continuation of The George Washington University Human Resources Research Office. HumRRO's general purpose is to improve human performance, particularly in organizational settings, through behavioral and social science research, development, and consultation.

Published
October 1970
by
HUMAN RESOURCES RESEARCH ORGANIZATION
300 North Washington Street
Alexandria, Virginia 22314

Prefatory Note

This paper is an updated documentation of a presentation made by the author, a Research Scientist at the Human Resources Research Organization Division No. 2, Fort Knox, Kentucky, at a THEMIS conference at Texas Technological University. The THEMIS contract is monitored by the U.S. Army Human Engineering Laboratories.

Mr. Engel also presented a paper, "An Approach to the Classification and Evaluation of Job Performance Measures," based on the same materials as in this publication, at the 12th annual conference of the Military Testing Association held at French Lick, Indiana, in September 1970.

The research reported in this presentation was conducted under HumRRO Work Unit JOBTEST, Proficiency Measurement Techniques.

AN APPROACH TO STANDARDIZING HUMAN PERFORMANCE ASSESSMENT

John D. Engel

I will discuss some factors I believe to be critical in standardizing and evaluating methods of performance assessment, concentrating on two primary factors—a task classification system and a performance measure, or criterion classification system. I also want to briefly describe the research program we at HumRRO have been conducting on the use of a performance measure classification system.

A TASK CLASSIFICATION SYSTEM

Let us first examine the factor of a task taxonomy. Gagne (1), speaking of the development of categories of human equipment operation behavior, stated that " . . . although a kind of taxonomy must probably be involved, the important research problem appears to be the development of a theoretical system which will relate physical task variables to performance variables by means of conceptualized intervening process. The lack of such a theory creates a void in this area of human behavioral knowledge." Essentially, Gagne was commenting on the absence of categories into which equipment operation tasks might be placed. These categories, if they existed, would have direct implications for standardizing performance measurement techniques.

There have been a number of attempts to solve this taxonomic problem. Cotterman (2) stated that " . . . a task classification scheme be developed in terms of which it is possible to sort all human learning tasks. Each task category would be set up in such a way that a specified set of common principles of learning referring to basic training variables would operate in essentially the same way in all task situations subsumed under it. In this way the actual and hypothesized effects of various basic [training] and task variables and their interactions would be set forth."

Although he uses many of the same distinctions as Cotterman, Stolurow (3) has proposed a systems approach to the development of a task taxonomy.

Haggard (4), in an extensive review of taxonomies, has suggested that as an organizational directive for constructing a classification system for psychological phenomena, we must undertake two distinct but related efforts: (a) build a theoretical structure that would provide the criterial priorities and definitions for a generalized conceptual system under which to interrelate psychological phenomena at all levels of generality and possibly to integrate psychological phenomena into the broader structure of the biological sciences; (b) deal only with the level of generality that is the primary concern of the training psychologist. The latter effort would provide a system of categories for

relating the principles of training to the definitions of behavior for that level of generality. The purpose would be to enable us to order information on training more coherently, so as to provide a means for interpreting and controlling training processes.

These efforts represent systematic and long-range approaches to the problem of classifying human performances in meaningful terms with respect both to the significance of learning principles and the importance of specific task influences. They are not yet suitable for practical application. According to Parker and Downs (5), at the present time only two practical methods of classifying human performance are available, classification by content and classification by performance.

The central theme of content classification methods is that certain inferences can be made concerning the cognitive or perceptual-motor skill requirements of a task. Schwarz (6), for example, discusses two broad types of knowledges required for effective job performance. One set consists of specific and unique items of information, for example, terminology, tolerance limits, task sequences. The second is more general in nature, for example, wiring procedures, computational routines, and mechanics of combustion agencies.

The central theme of performance classification methods is to express task activities in terms of the type of performance involved. For example, Schwarz (6) presents the following list which typifies use of this type of classification scheme:

(1) Visual discrimination

(2) Auditory discrimination

(3) Manipulation

(4) Decision making

(5) Symbolic data operation

(6) Reporting

We have briefly covered representative ideas in the task taxonomy area; the main purpose in this paper is to consider data concerning the relationship of a task taxonomy and performance measurement taxonomy. I do not intend to develop or propose a new theoretical model to relate task classification and performance measurement. However, some classification scheme is required for the approach we are proposing and, therefore, it will be necessary to select one already available.

Before discussing the classification scheme that was selected, another consideration should be mentioned. Miller (7) has stated that there are various purposes for which one might formulate a task taxonomy: (a) Predicting the skill level of various trainees on particular tasks, such as those in selection tests, factor analytic studies, or simple correlation studies; (b) designing equipment so that particular tasks may be performed more efficiently; (c) determining which training strategies or educational techniques are most appropriate for particular tasks; (d) discerning which underlying learning processes are the most important ones in the acquisition of particular tasks.

Certainly there is no one appropriate taxonomy for all these purposes. Thus in our case, a taxonomy that would enable us to determine which performance measurement strategy is most appropriate for a particular task or category of tasks is needed.

Parker and Downs (5) have suggested using a system developed by Lumsdaine (8) and subsequently modified by them to be useful in training studies [which include performance measurement]. The authors cite two reasons for selecting this particular scheme. They are: (a) it appears, after using the system to classify system activities of an Air Force Tactical Air Control System, that items of human performance data can be assigned into the appropriate category, (b) assignment to a given category implies that the specific type of training [also measurement] will be more appropriate for the training of the task than will other types. I feel this second reason is rather tenuous and must be empirically established. In any case, the six classes in Lumsdaine's system are:

(1) Learning Identifications. This means pointing to or locating objects and locations, naming them or identifying what goes with what.

(2) Perceptual Discriminations. This involves the use of visual, auditory and similar cues in a manner which allows the identification of a particular stimulus.

(3) Principles' and Relationships' Comprehension. This involves understanding a statement of relationship as evidenced by being able to state, illustrate and recognize its implications.

(4) Procedural Sequencing. This means carrying out a set of operations that must be performed in a fixed sequence.

(5) Decision Making. This involves the application of conceptual rules or principles as the basis for making the kinds of decisions that are involved in diagnosing or interpreting complex situations.

(6) Perceptual-Motor Skills. These may be simple, such as using a basic hand tool, or complex, such as manipulating the controls of an airplane.

It should be noted that these categories are not mutually exclusive. Parker and Downs (5) give a good example when they state:

> For example, a maintenance duty such as *align equipment components* obviously involves two responses. Initially, the maintenance man must be trained in following procedural sequence. There is a set pattern of alignment procedures which is appropriate to this equipment complex. However, during the course of following these procedures, there are essential motor skills which are required in bringing each separate adjustment into tolerance. For training [or testing] purposes it is important that such an activity be classified within both categories. The training [or testing] must account for the *following procedures* portion of the activity as well as the perceptual-motor skill portion.

We have what appears to be a workable, although not ideal classification system. However, two things must be emphasized: (a) This taxonomy is used for illustrative purposes only, and (b) I have discussed a task taxonomy for training purposes and have assumed that this type of taxonomy would also be relevant for performance measurement purposes. This may not be true, and we will, therefore, have to make adjustments in the taxonomy as our research efforts proceed.

A PERFORMANCE MEASURE CLASSIFICATION SYSTEM

The second factor we are considering is that of a classification system for performance measures.

There are various types of proficiency measurement techniques that might be used to evaluate a man's achievement. According to Glaser and Klaus (9), proficiency measurement techniques may be grossly categorized on the basis of their remoteness from actual job performance. This remoteness may be due to differences in (a) the behavior elicited for measurement, (b) the eliciting stimuli themselves, or (c) both stimulus and behavior. In most instances, however, as the test stimuli become more remote from those found in the actual job situation, the responses elicited are likewise less similar to those found in job performance.

Thus at one extreme along this continuum of remoteness is the measurement of proficiency during actual job performance. At the other extreme are measures (e.g., paper-and-pencil tests) that are not obviously similar to the criterion task, but assess performance on tasks that correlate with on-the-job behavior.

Between these two extremes are test situations that: (a) call for the performance of the actual job task outside the real job environment, or (b) attempt to simulate the job task while at the same time offering effective control of the factors that in "real" situations are likely to interfere with reliable and valid measurement. The four major segments along this continuum can be identified as (a) on-the-job measures, (b) work sample measures, (c) simulated-job measures, and (d) correlated-job measures.

In principle, proficiency measurement should be accomplished during a man's typical performance, under conditions generally present during day-to-day operations. This method, however, presents a number of problems. The degree of control that can be achieved in a job situation is generally less than satisfactory for obtaining reliable measurements. In addition, attempts to standardize the situation for proficiency-measurement purposes frequently introduce considerable artificiality into the situation. Finally, the consideration of committing large amounts of time, money, and men to the testing situation often makes this an impractical method of assessment.

To reduce, to some extent, the problems involved in on-the-job measures, samples of the actual job tasks involved may be removed from the real job environment so they can be readily and reliably assessed. This type of proficiency measurement technique is referred to as a "work sample test." Here, the individual performs the actual tasks but not in the real job environment. This technique is a close

approximation to on-the-job measures, but it has some of the same "drawbacks"—it is costly, time consuming, and essentially impractical as a method of assessing large numbers of people.

Because it is difficult to measure men's proficiency during actual job and work sample situations, the job may be simulated in a controlled manner in order to produce a reliable and valid, yet practical method of performance assessment. The essence of task simulation is the design of test stimuli that will evoke job-like responses that can be measured objectively. This general category of simulated-job measures includes a variety of proficiency measurement techniques. Some of the most frequently employed measures use equipment mock-ups and simulators.

An extreme position along the dimension of remoteness from job reality is represented by tests measuring, not job behaviors themselves, but correlated-job behaviors—that is, measures correlated with job behavior. These measures are the most remote from the actual job situation. The most widespread type of correlated-job measure is verbal response as used to assess skills that are substantially nonverbal. Examples of this type of proficiency measure are tests of job knowledge, vocabulary, and nomenclature used to evaluate performance at procedural and manipulative tasks. Other types of correlated-job measures are those that involve a deliberate modification in the response made so as to facilitate the recording and evaluation of responses. A common example of this kind of construction is multiple-choice paper-and-pencil tests that are used to measure the ability to produce appropriate responses by measuring ability to recognize them. Because they are easily constructed, inexpensive, and easily administered, paper-and-pencil tests of job knowledge are frequently used to evaluate an individual's proficiency.

However, tests measuring knowledge of technical information, tool nomenclature, technical vocabulary, or underlying theory may not relate to actual performance for some tasks. Instead, they measure verbal knowledge about the job, and therefore assess behaviors which, at best, may be correlated only slightly with actual job behavior—especially if the job depends on motor and manipulative skills.

Thus we now have the beginning of a performance measurement classification system—the second necessary factor in our approach to measurement standardization.

SOME RESEARCH RESULTS

The relationship between the two taxonomies is illustrated in Figure 1.

One approach to standardizing performance measures would be to determine empirically the validity of each of the performance measurement categories for each of the task categories. The data from this type of comparison would provide valuable information for making decisions concerning the most effective type of measurement to use for a particular type of task. This approach can be illustrated by using some data from HumRRO Work Unit JOBTEST.

Relationship Between Task and Performance Measurement Classification Systems

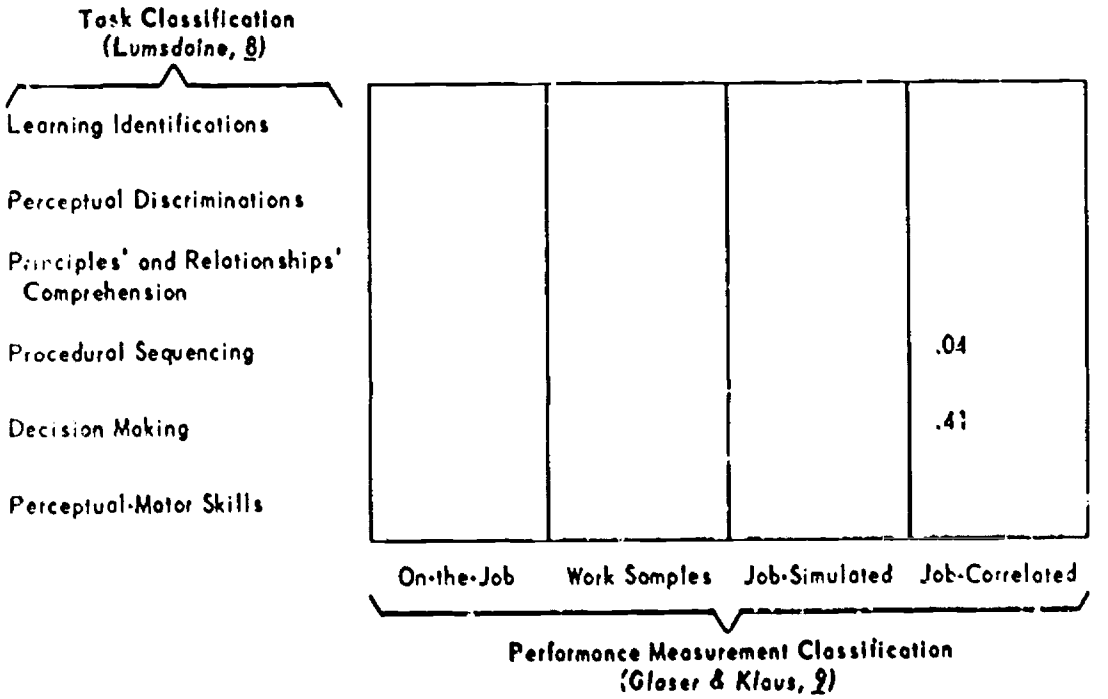

Figure 1

The major long-term research problem in JOBTEST was to study and evaluate a variety of concepts and procedures for the measurement of job performances. Emphasis was placed on identifying those techniques that have both validity and utility in practical testing environments, and that have generality across groups of tasks.

The first phase of the research was the development of a relevant and reliable work sample criterion for the General Vehicle Mechanic. This criterion was used as a standard in later research phases that evaluated various measurement techniques.

Work was begun by updating job information in a 1964 HumRRO analysis of job requirements for consolidated MOS 630, 631, 632 (Automotive Mechanic). This inventory was used as a basis for developing items for a "hands-on-equipment" work sample.

A four-day proficiency test consisting of 33 sample exercises was constructed. The test included a diagnostic scoring procedure for use in scoring men on quality of performance. The exercises were individually performed on track and wheel vehicles in common use and were individually scored by experienced mechanics who had been trained in proper test administration procedures.

The test was administered to 38 organizational mechanics, drawn from all organizational maintenance units at Fort Knox, Kentucky. In

addition, a questionnaire was used to obtain information on personnel data, organizational maintenance experience, experience on various vehicle systems, current job assignment, type and amount of training, and amount of supervision received on the job.

The results indicated that the total test appears to have a high degree of reliability (r = .82), indicating it should permit a high degree of accuracy of measurement when used as a criterion in evaluating other measurement techniques.

The second phase of research dealt with a comparison of two job-correlated measures with the work sample.

Work was begun by recalling 30 organizational mechanics who had been subjects during the earlier development of the work sample criterion for general vehicle repairman.

The 30 organizational mechanics were given the appropriate paper-and-pencil MOS Evaluation Test as developed by the Enlisted Evaluation Center. Approximately one week following the administration of the written test, peer ratings were collected on each subject in accordance with procedures established by the Enlisted Evaluation Center. The results of this work indicated:

(1) When correlated with the work sample criterion, the written test was shown to have a low degree of validity (r = .27); this value is too low for use of the test in group or individual measurement.

(2) When correlated with the work-sample criterion, the peer ratings were shown to have a low degree of validity (r = .24); this value is too low for use of the ratings in group or individual measurement.

(3) There was an extremely low relationship between the written test and the peer ratings (r = .06); too low for use in group or individual measurement.

(4) When correlated with the work-sample criterion, the troubleshooting items on the written test were shown to have a moderate degree of validity (r = .41); this value is high enough for such items to be useful for group measurement.

(5) When correlated with the work-sample criterion, the corrective action items on the written test were shown to have a low degree of validity (r = .04); too low for use in group or individual measurement.

If these results are entered in Figure 1 with the troubleshooting tasks coded as essentially "decision making" tasks and corrective action tasks as essentially "procedural sequencing" tasks, we see that the data indicated that job-correlated measures are more valid indicators of performance on decision-making tasks than on procedural sequencing tasks.

One reservation regarding the preceding data and approach should be mentioned. First, we have assumed that if we were to place troubleshooting tasks into one, and only one, of Lumsdaine's categories, it would have to be put in the "decision-making" category. In reality,

troubleshooting activity is composed of many tasks and these would have to be determined and probably differentially weighted across various categories in Lumsdaine's system. The nature of our problem is certainly not as simple or clear-cut as it is portrayed in the example, which is purposely simple to illustrate the approach.

AREAS FOR FUTURE RESEARCH

I believe there is a need for research along at least two parallel lines:

(1) The development and refinement of an interim task classification system along the lines suggested by Haggard (4). To reiterate, Haggard holds that: "The particular degree of generality which is most applicable to the activities of the training psychologist is the degree which focuses on the complex knowledges and skills determined by the systems analysis, not the one which is implied by the traditional learning situation. Analysis at the level of systems analysis should supply a structure of essential categories which are intrinsically interrelated at that degree of generality."

(2) The development and refinement of an interim classification system for human performance measures along the lines suggested by Glaser and Klaus (9) and the validation of such measures using currently available data about task dimensions or categories. One approach has been described in this paper along with some preliminary data from Work Unit JOBTEST.

Finally, the interaction of these two research approaches will probably yield a "mixed measurement technique" which advocates the use of different measurement techniques for different types of tasks within the total job.

LITERATURE CITED

1. Gagne, Robert M. "Training Devices and Simulators: Some Research Issues," *The American Psychologist*, vol. 9, no. 3, March 1954, pp. 95-107.

2. Cotterman, T. E. *Task Classification: An Approach to Partially Ordering Information on Human Learning*, WADC Technical Note 58-374, Wright Air Development Center, Wright-Patterson AFB, Ohio, January 1959.

3. Stolurow, Lawrence M. *A Taxonomy of Learning Task Characteristics*, Technical Documentary Report No. AMRL-TDR-64-2, Behavioral Sciences Laboratory, Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command, Wright-Patterson AFB, Ohio, January 1964.

4. Haggard, Donald F. *The Feasibility of Developing a Task Classification Structure for Ordering Training Principles and Training Content*, HumRRO Research Memorandum, January 1963.

5. Parker, J. F., and Downs, J. E. *Selection of Training Media*, ASD-TR-61-473, Behavioral Sciences Laboratory, Aerospace Medical Laboraotry, Aeronautical Systems Division, Air Force Systems Command, USAF, Wright-Patterson AFB, Ohio, September 1961.

6. Schwarz, P. A. "Design of Selection and Training Procedures," in *Human Factors Methods for System Design*, John D. Folley, Jr. (ed.), Chapter 10, AIR-290-60-FR-225, American Institute for Research, Pittsburg, Pennsylvania, 1960.

7. Miller, E. E. *A Taxonomy of Response Processes*, HumRRO Technical Report 69-16, September 1969.

8. Lumsdaine, Arthur A. "Design of Training Aids and Devices," in *Human Factors Methods for System Design*, John D. Folley, Jr. (ed.), Chapter 11, AIR 290-60-FR-225, American Institute for Research, Pittsburg, 1960.

9. Glaser, Robert, and Klaus, David J. "Proficiency Measurement: Assessing Human Performance," in *Psychological Principles in System Development*, R. M. Gagne (ed.), Holt, Rinehart and Winston, Inc., New York, 1962.

Unclassified
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Human Resources Research Organization (HumRRO) 300 North Washington Street Alexandria, Virginia 22314	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP

3. REPORT TITLE

AN APPROACH TO STANDARDIZING HUMAN PERFORMANCE ASSESSMENT

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Professional Paper

5. AUTHOR(S) (First name, middle initial, last name)

John D. Engel

6. REPORT DATE October 1970	7a. TOTAL NO. OF PAGES 12	7b. NO. OF REFS 9
8a. CONTRACT OR GRANT NO. DAHC 19-70-C-0012 b. PROJECT NO. 2Q062107A712 c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) Professional Paper 26-70 9b. OTHER REPORT NO.(S) (Any other numbers that may be assigned this report)	

10. DISTRIBUTION STATEMENT

Distribution of this document is unlimited.

11. SUPPLEMENTARY NOTES Presented at Conference on Human Factors Research, Texas Technological University, Lubbock, March 1970	12. SPONSORING MILITARY ACTIVITY Office, Chief of Research and Development Department of the Army Washington, D.C. 20310

13. ABSTRACT

The standardization and evaluation of methods of performance assessment represents an important area of concern. In this paper an approach that concentrates on two critical areas and the relationship between them is discussed. These are: (a) a task classification system, and (b) a performance measure classification system. An example is presented that illustrates some preliminary research related to the use of a performance measure classification system. The paper concludes by suggesting areas and directions for future research efforts.

DD FORM 1 NOV 68 1473

Unclassified
Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Criterion Development						
Human Performance						
Job-Correlated Measures						
Performance Assessment						
Proficiency						
Task Classification						
Work-Sample Criterion						